SNR Classification System Based on Classification of Voiced/Unvoiced Signal

Jae Seung Choi [1]

[1] Department of Electronic Engineering, College of Engineering, Silla University, 140 Baegyang-daero (Blvd), 700 Beon-gil (Rd), Sasang-gu, Busan, 617-736, Korea
jschoi@silla.ac.kr

Abstract. This paper proposes a signal-to-noise ratio (SNR) classification system based on a classification of voiced/unvoiced signal using a time-delay neural network for noise reduction in speech that is degraded by background noises. As such, the proposed system detects voiced and unvoiced sections, then reduces the noise signal for each input frame using the time-delay neural network.

Keywords: Classification system, signal-to-noise ratio, time-delay neural network, background noise.

1 Introduction

For speech signal processing, a neural network (NN) needs to be constructed using a time structure, as the time variation is significant information [1, 2]. Moreover, an amplitude component contains more information than a phase component when a speech signal is generated by a fast Fourier transform (FFT). Accordingly, this paper proposes a signal-to-noise ratio (SNR) classification system based on a classification of voiced/unvoiced signal using a time-delay neural network (TDNN) [3, 4, 5, 6], which includes a time structure in the NN, then confirms the efficiency of the proposed system based on experiments of classification rates in a speech signal. Using the correct classification rates, experiments confirm that the proposed system is effective for speech degraded by noises, such as white and subway noise.

The remainder of this paper is organized as follows. Section 2 introduces the construction of the proposed time-delay neural network. Section 3 explains a speech and noise database used in the experiments and discusses the experimental results when using the proposed system. Section 4 presents some conclusions.

2 Proposed time-delay neural network (TDNN)

Fig. 1 shows the construction of the proposed TDNN system. First, the noisy speech signal $x(k)$ is divided into length frames of 128 samples (16 ms). Next, the $x(k)$ is detected in the voiced and unvoiced sections, then separated into FFT amplitude components according to the voiced and unvoiced sections. Thereafter, the separated

© SERSC 2013

FFT amplitude components are added to the appropriate TDNNs with the low, mid and high frequency bands.

In this experiment, the TDNNs for the low, mid, and high frequency band are composed of four layers and the compositions of the TDNNs are 22-60-22-22. The final FFT amplitude component is then obtained by combining the results from the TDNNs with the low, mid and high frequency bands. However, the FFT phase component is directly obtained from the original noisy speech signal, after detecting voiced and unvoiced sections. Thereafter, the enhanced speech signal $y(k)$ is regenerated using an inverse fast Fourier transform (IFFT).

In the training of the proposed TDNNs, the training coefficient was set to 0.2 and the inertia coefficient was set to 0.7. Moreover, random numbers from –1.0 to 1.0 is used as an initial weight, and 10,000 was set as the maximum number of training iteration for the experiment.

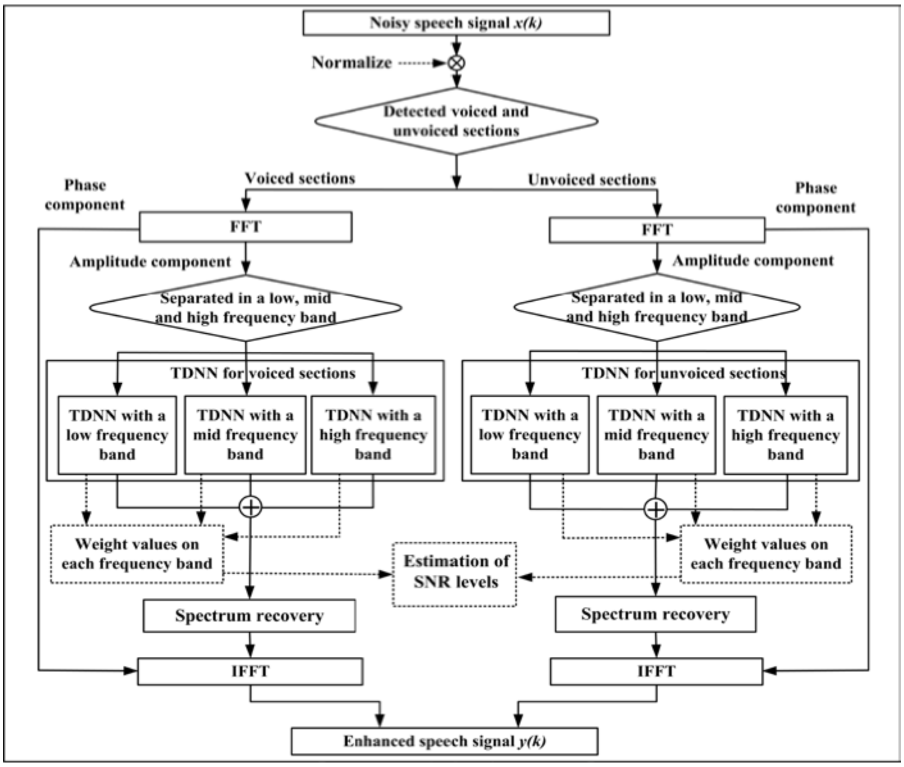
Noisy speech signal x(k)
Normalize
Detected voiced and unvoiced sections
Phase component
Voiced sections
Unvoiced sections
Phase component
FFT
FFT
Amplitude component
Amplitude component
Separated in a low, mid and high frequency band
Separated in a low, mid and high frequency band
TDNN for voiced sections
TDNN for unvoiced sections
TDNN with a low frequency band
TDNN with a mid frequency band
TDNN with a high frequency band
TDNN with a low frequency band
TDNN with a mid frequency band
TDNN with a high frequency band
Weight values on each frequency band
Estimation of SNR levels
Weight values on each frequency band
Spectrum recovery
Spectrum recovery
IFFT
IFFT
Enhanced speech signal y(k)

Fig. 1. The construction of the proposed TDNN system.

3. Experimental results

All speech data of the Aurora-2 database consists of English-connected digits spoken by American English speakers [7, 8, 9]. In this experiment, the proposed system was evaluated using speech data from the Aurora-2 database in Test Sets A, B, and C and

SNR Classification System Based on 125 Classification of Voiced/Unvoiced Signal

two types of background noise, i.e. subway noise in Test Set A, and white noise generated by a computer program. In this experiment, the TDNNs are trained using noisy speech data artificially added at several SNRs (20 dB, 15 dB, 10 dB, 5 dB, and 0 dB). When using the Aurora-2 database, the TDNNs are trained after adding white and subway noise to the clean speech data in the Aurora-2 database.

The performance of the proposed TDNN system was tested based on the correct classification rate, frame-by-frame, and the definition of the classification rate was the ratio of the number of frames in which the SNR levels were correctly estimated to the total number of frames given as the input.

Fig. 2 shows the average values of the classification rates of the proposed TDNN system for the noises, when using a total of twenty different test utterances selected from Test Sets A and B. In the case of TDNN with the low frequency band when unvoiced sections, the classification rates averaged over fifty utterances were 87% or more for each condition of white and subway noises in Test Sets A and B. However, the average values of the classification rates were approximately 3% worse for such noises, in the case of TDNN with the mid and high frequency bands when voiced and unvoiced sections, respectively.

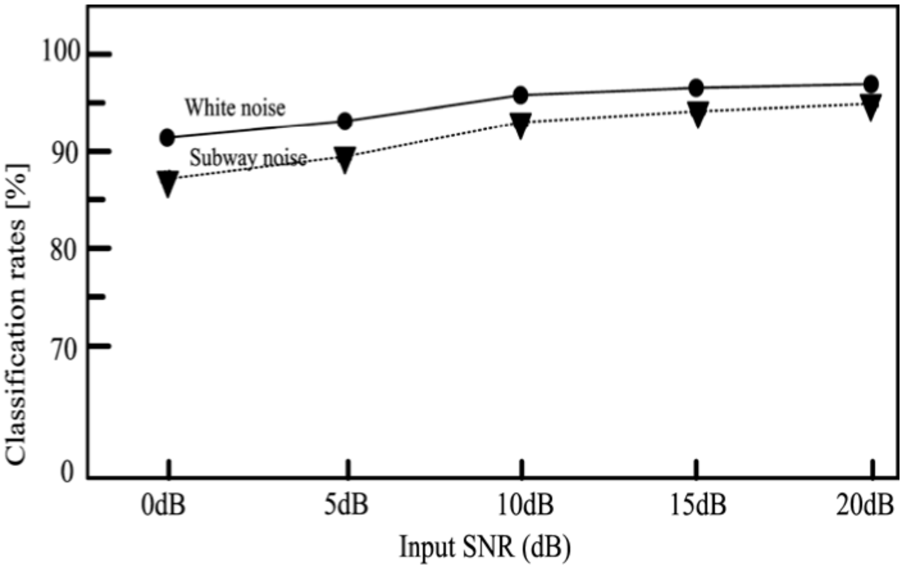

Fig. 2. Classification rates for TDNN with the low frequency band when unvoiced sections.

4 Conclusions

A TDNN system based on classification of voiced/unvoiced signal was proposed that uses TDNN to reduce background noise. Experimental results confirmed that the proposed system is effective for white and subway noise, as demonstrated by the classification rates. In summary, the experimental results were as follows:

1. The possibility of noise classification using a TDNN was confirmed in this experiment.
2. The noise reduction was significant under input SNR conditions of up to about 0 dB for sentences.

The following problems remain as future areas for study.

1. The effectiveness of the proposed system needs to be evaluated for speech degraded by heavy noise and various non-stationary noises in a real environment.

As mentioned above, the proposed system using the TDNN was experimentally demonstrated for white and subway noise. Therefore, it is believed that the present research results will be useful for the speech recognition under noisy conditions.

References

1. S. Furui, S., D. Itoh: Neural-network-based HMM adaptation for noisy speech, IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 365-368, (2001)
2. K. Daqrouq, I.N. Abu-Isbeih, M. Alfauri: Speech signal enhancement using neural network and wavelet transform, Proceedings of the 6th International Multi-Conference on Systems, Signals and Devices, Djerba, Tunisia, pp. 1-6, (2009)
3. M. Debyeche, A. Amrouche, J.P. Haton: Distributed TDNN-Fuzzy Vector Quantization For HMM Speech Recognition**,** International Conference on Multimedia Computing and Systems, pp. 72-76, April (2009)
4. R.A. Mitchell, A. Shaw: Vowel recognition with a time-delay neural network, IEEE International Conference on Systems Engineering, pp. 637-640, (1990)
5. J.B. Hampshire, A.H. Waibel: A novel objective function for improved phoneme recognition using time delay neural networks, IEEE Transactions on Neural Networks, vol. 1, no. 2, pp. 216-228, (1990)
6. J.S. Choi, S.J. Park: Speech Enhancement System based on Auditory System and Time-Delay Neural Network, 8th International Conference on Lecture Notes in Computer Science, Part II, pp. 153-160, April (2007)
7. H. Hirsch, D. Pearce: The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions, in Proc. ISCA ITRW ASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, (2000)
8. Leonard R.G.: A database for speaker independent digit recognition, IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 328-331, Mar (1984)
9. ITU-T (International Telecommunication Union) recommendation G. 712: Transmission performance characteristics of pulse code modulation channels, pp. 1-31, (1996)